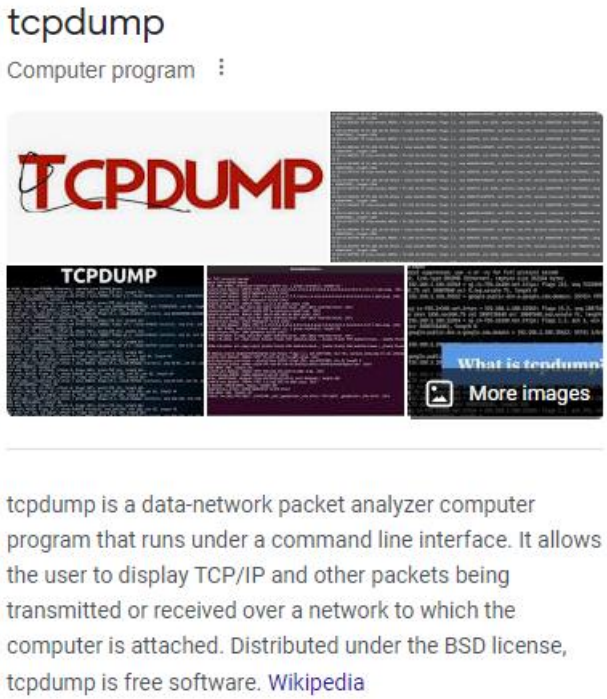
Project Pro 50 - Final Project

Intrusion Detection System - UNSW_NB15
Streaming - Processing - Analysis

Business Overview:

The raw network packets of the UNSW-NB 15 dataset was created by the IXIA Perfect Storm tool in the Cyber Range Lab of the Australian Centre for Cyber Security (ACCS) for generating a hybrid of real modern normal activities and synthetic contemporary attack behaviours.



Tcp dump tool is utilised to capture 100 GB of the raw traffic (e.g., Pcap files). This dataset has nine types of attacks, namely, Fuzzers, Analysis, Backdoors, DoS, Exploits, Generic, Reconnaissance, Shellcode and Worms. The Argus, Bro-IDS tools are used and twelve algorithms are developed to generate totally 49 features with the class label. The Challenge here is there are multiple Algorithms generating lots of data with 49 features and it is a critical application and has to be taken action in Real-Time. But due to the volume of data it is difficult to analyse the data in real-time and take timely action. So we need to build a robust data pipeline that can read data in real-time and provide a dashboard where we can analyse any alerts and take immediate action.

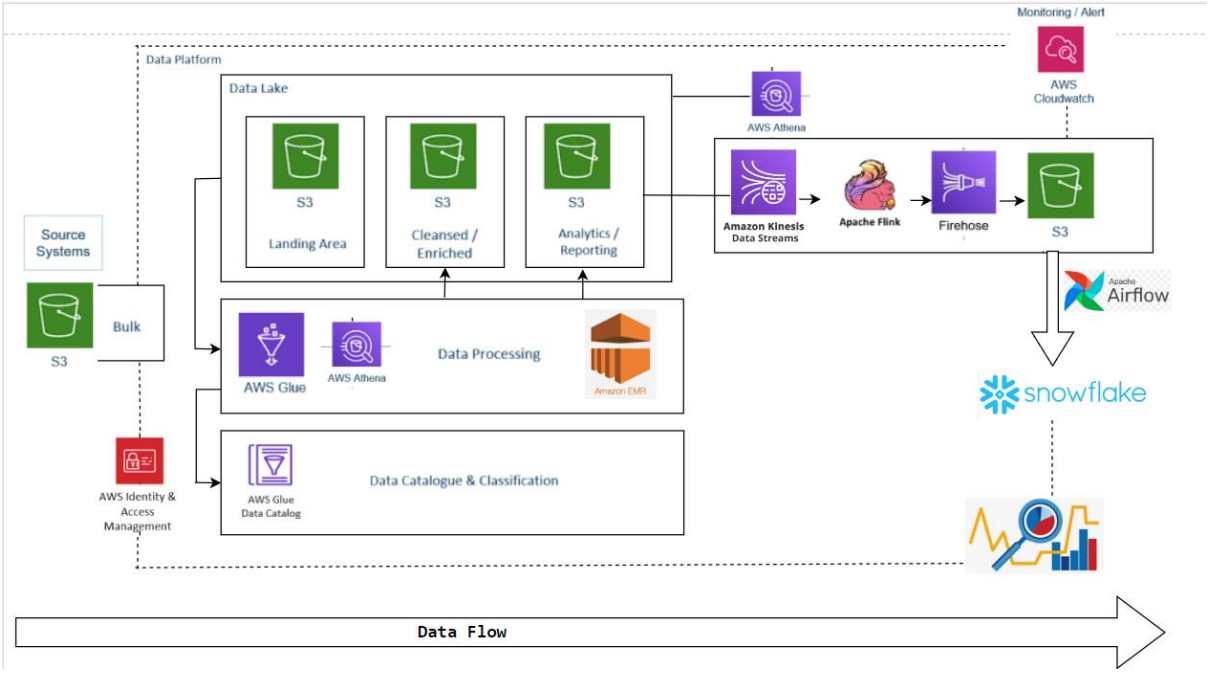
Data Pipeline:

So we will be receiving the data from the API and receive it in form of updated csv file in AWS S3 bucket. Once the Data lands in S3 we will have to Clean the data by doing some filtrations and we will achieve this using AWS Glue and Aws Lambda and store the transformed data in S3 bucket. Now we will Process the data further in AWS EMR and again stored in a new AWS S3 bucket. The data that is transformed is now Validated through AWS Athena.

Now this is happening in Live so we will have to stream the data to a dashboarding tool and we will achieve this through AWS Kinesis Data Stream from where the data will be read through Kinesis Firehouse that could update Snowflake Tables. The whole pipeline of Streaming is automated using Airflow.

From Snowflake data can be connected to dashboarding tools and visualized in real time. The issues that might arise here is Permissions to handle all the AWS Data Services and concerns with the scripts that we can trouble shoot with AWS cloud watch. The Data Flow pipeline architecture is drafted below.

Architecture:



Dataset Description:

The raw network packets of the UNSW-NB 15 dataset was created by the IXIA PerfectStorm tool in the Cyber Range Lab of the Australian Centre for Cyber Security (ACCS) for generating a hybrid of real modern normal activities and synthetic contemporary attack behaviours. There are a total of 49 Feature columns where we will be carrying with the Data Pipeline for analysis.

Tech Stack:

- ➔ **Languages:** Python, SQL, SPARK,
- ➔ **AWS Services**: AWS S3, AWS Glue, AWS Athena, AWS EMR, Apache Flink, Amazon Kinesis, Amazon IAM, AWS Lambda, Amazon CloudWatch, Apache Zepplin, Apache Airflow,
- ➔ **Other services and tools:** Snowflake, VS code

1. **Python**: Python is a versatile, high-level programming language known for its readability and simplicity. It supports multiple programming paradigms and is widely used in web development, data analysis, artificial intelligence, and more.

2. **SQL**: SQL (Structured Query Language) is a domain-specific language used for managing and querying relational databases. It allows users to interact with databases to retrieve, update, and manipulate data in a structured format.

3. **SPARK**: Apache Spark is an open-source distributed computing system that provides a fast and general-purpose cluster-computing framework. It's used for big data processing and analytics tasks, offering high performance and fault tolerance.

4. **AWS S3**: Amazon Simple Storage Service (S3) is an object storage service provided by Amazon Web Services. It offers scalable, secure, and reliable storage for data of various types, making it widely used for data backup, archiving, and web hosting.

5. **AWS Glue**: AWS Glue is a managed extract, transform, and load (ETL) service provided by Amazon Web Services. It simplifies the process of preparing and loading data for analytics by automatically discovering, cataloging, and transforming data.

6. **AWS Athena**: Amazon Athena is an interactive query service that allows you to analyze data stored in AWS S3 using standard SQL queries. It provides a serverless, pay-per-query approach to data analysis.

7. **AWS EMR**: Amazon Elastic MapReduce (EMR) is a cloud-based big data platform provided by AWS. It simplifies the processing of large datasets by offering a managed Hadoop framework, along with other big data processing tools.

8. **Apache Flink**: Apache Flink is an open-source stream processing and batch processing framework. It's designed for high-throughput, fault-tolerant, and low-latency processing of streaming data.

9. **Amazon Kinesis**: Amazon Kinesis is a managed real-time data streaming service provided by AWS. It allows you to collect, process, and analyze large streams of data in real-time.

10. **Amazon IAM**: Amazon Identity and Access Management (IAM) is a service that allows you to manage user accounts and permissions for accessing AWS services and resources. It helps control who can do what within an AWS account.

11. **AWS Lambda**: AWS Lambda is a serverless compute service provided by AWS. It allows you to run code in response to events without having to provision or manage servers, making it ideal for event-driven applications.

12. **Amazon CloudWatch**: Amazon CloudWatch is a monitoring and observability service provided by AWS. It allows you to collect and track metrics, collect and monitor log files, and set alarms for your AWS resources.

13. **Apache Zeppelin:** Apache Zeppelin is an open-source web-based notebook for data analytics, data ingestion, data exploration, and data visualization. It provides an interactive environment for data scientists and analysts.

14. **Apache Airflow**: Apache Airflow is an open-source platform used for programmatically authoring, scheduling, and monitoring workflows. It allows you to define complex workflows as code.

15. **Snowflake**: Snowflake is a cloud-based data warehousing platform that allows businesses to store and analyze large amounts of data in a scalable and efficient manner. It's known for its performance, simplicity, and versatility.

16. **VS Code**: Visual Studio Code (VS Code) is a free, open-source code editor developed by Microsoft. It supports various programming languages and offers features like debugging, syntax highlighting, version control, and extensions for customization.

Key Takeaways:

- Understanding the project Overview and Architecture
- Understanding ETL on Big Data
- Introduction to Staging and Data Lake
- Creating IAM Roles and Policies
- Understanding the Dataset
- Handling AWS GLUE, Crawlers
- Scripting Glue jobs and troubleshooting them
- Writing Pyspark code on Glue
- Querying in Athena
- Spinning AWS EMR clusters
- Accessing EC2 via SSH
- Setting up AWS CLI
- Implementing Pyspark code from EMR clusters
- Understanding Data Streams and Amazon Kinesis
- Understanding Apache Flink
- Understanding Kinesis Firehose
- Creating a Kinesis Data Analytics Application
- Working with Apache Zepplin Notebooks
- Using CloudWacth to monitor logs and metrics
- Understanding Snowflake and its significance in the modern Datawarehouse ecosystem
- Setting up Snowflake and establishing secure connection with AWS
- Understanding Airflow
- Creating Dags and Data orchestration
- Confidence to create any data pipeline with any scale and understanding the Business impact of it and how important the Architecture is to the Data world.